Tree Densities

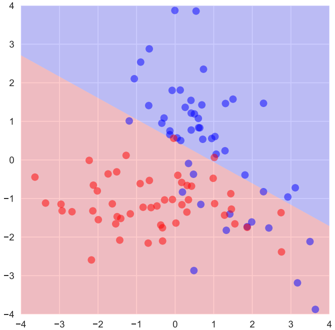
Nick Landolfi and Sanjay Lall
Stanford University

Outline

Motivation

Motivation: classification

- we have a *dataset* of *records* $u^1, \dots, u^n \in \mathcal{U}$ and $v^1, \dots, v^n \in \mathcal{V}$ with $\mathcal{V}$ a finite set of *classes*
- we want to build a *classifier* $G : \mathcal{U} \to \mathcal{V}$ and use it to classify a new *independent variable* u as $G(u)$
- for example, $\mathcal{U} = \mathbf{R}^2$ and $\mathcal{V} = \{0, 1\}$



  - the point u^k is colored red if $v^k = 0$ and blue if $v^k = 1$
  - the region $\left\{u \in \mathbf{R}^2 \;\middle|\; G(u) = 0\right\}$ is shaded red and $\left\{u \in \mathbf{R}^2 \;\middle|\; G(u) = 1\right\}$ is shaded blue

Example application

- for example, $\mathcal{U} = \mathbf{R}^{1000}$ and $\mathcal{V} = \{0, 1\}$
  - u represents the expression signature of 1000 genes
  - v indicates the condition of mutation of a particular gene, upon which prognosis varies

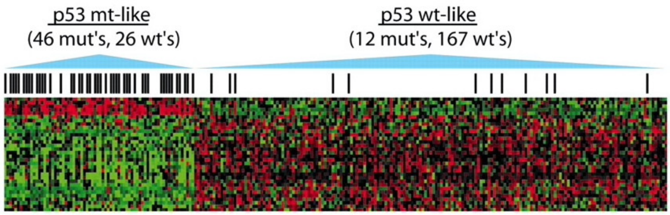


Figure 1: visualization of gene expression data

- one approach is to produce a density over $\mathcal{U}$ for each class; called *generative* modeling
  - for a new u, we define $G(u)$ to be a class with *maximum likelihood*
  - binary classification with normals: *linear* or *quadratic discriminant analysis*

Tree densities

Densities

- a *density* is a function $f : \mathbf{R}^d \to \mathbf{R}$ with $f \geq 0$ and $\int f = 1$
- the *ith marginal density* of f is $f_i : \mathbf{R} \to \mathbf{R}$ so that

$$f_i(\xi) = \int_{x_i = \xi} f(x)dx$$

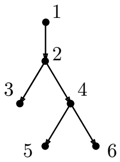
  for all $\xi \in \mathbf{R}$, for $i = 1, \ldots, d$
  - similarly, $f_{ij}(\xi, \gamma) = \int_{x_i = \xi, x_j = \gamma} f(x)$
- the *i, jth conditional* of f is a function $f_{i|j} : \mathbf{R}^2 \to \mathbf{R}$ satisfying

$$f_{ij}(\xi, \gamma) = f_{i|j}(\xi, \gamma) f_j(\gamma)$$

  for $\xi, \gamma \in \mathbf{R}$, for $i, j = 1, \ldots, d$ and $i \neq j$
- our simpler densities will be products of these one-variable marginals and two-variable conditionals

Tree density: example

- consider $T = \{\{1,2\},\{2,3\},\{2,4\},\{4,5\},\{4,6\}\}$ rooted at vertex 1



- if f is a density on $\mathbf{R}^6$, then by *chain rule* f always satisfies

$$f = f_{6|1,2,3,4,5}\, f_{5|1,2,3,4}\, f_{4|1,2,3}\, f_{3|1,2}\, f_{2|1}\, f_1$$

- we say f *factors* according to the tree T rooted at vertex 1 if f satisfies

$$f = f_{6|4}\, f_{5|4}\, f_{4|2}\, f_{3|2}\, f_{2|1}\, f_1$$

  - so $f_{6|1,2,3,4,5} = f_{6|4}$ (the conditional distribution does not depend on x_1, x_2, x_3 or x_5)
  - and similarly for $f_{5|4}$, $f_{4|2}$ and $f_{3|2}$

Tree densities

- **Definition:** a density f on $\mathbf{R}^d$ factors according to a tree T on $\{1, \ldots, d\}$ if it factors according to T rooted at some vertex
  - there exists i such that
$$f(x) = f_i(x_i) \prod_{j \neq i} f_{j|\mathsf{pa}_j}(x_j, x_{\mathsf{pa}_j})$$
  - if true for one vertex, true for all
- a density f need not factor according to a tree; may factor according to many trees
- we care because it may require fewer parameters to specify a tree density
  - for example, to specify a normal requires $O(d^2)$, if tree-structured then $O(d)$.

Tree density approximators

Differential Kullback-Leibler divergence

- we want a criterion for how well a density f approximates a density g
- we will use the *Kullback-Leibler divergence*, defined by

$$d_{kl}(g, f) = h(g, f) - h(g)$$

  - where $h(g) = -\int_{\text{supp}(g)} g(x) \log g(x) dx$ is called the *differential entropy*
  - and $h(g, f) = -\int_{\text{supp}(g)} g(x) \log f(x) dx$ is called the *differential cross entropy*
- we interpret d_{kl} as a measure of the difference between two densities
  - $d_{kl}(g, f) \geq 0$ for all densities g and $d_{kl}(g, g) = 0$.
  - if we want to find a density f to

$$\text{minimize} \quad d_{kl}(g, f)$$

    then $f = g$ is a solution; later we constrain f
  - d_{kl} is not symmetric and so not a metric, though we do not mind

Formulation

- have density g on $\mathbf{R}^d$
- want to find density f on $\mathbf{R}^d$ and tree T on $\{1, \dots, d\}$ to

$$\begin{array}{rl} \text{minimize} & d_{kl}(g, f) \\ \text{subject to} & f \text{ factors according to } T \end{array}$$

  - called the *tree density approximation* of g
  - call a solution pair an *optimal tree density approximator* and *optimal approximator tree* of g

Tree density approximator: first theorem

- **Theorem 1**: Let g be a density on $\mathbf{R}^d$. Let T be a tree on $\{1, \ldots, d\}$. Let $\mathsf{pa}_{(\cdot)}$ be defined by T rooted at vertex i. Then the density f_T^* on $\mathbf{R}^d$ defined by

$$f_T^* = g_i \prod_{j \neq i} g_{j \,|\, \mathsf{pa}_j}$$

achieves minimum Kullback-Leibler divergence to g among all distributions which factor according to T.

Proof of theorem 1

- we express the differential cross entropy of f relative to g by

$$
\begin{aligned}
h(g, f) &= -\int_{\mathbf{R}^d} g \log f \\
&= -\int_{\mathbf{R}^d} g(x) \left(\log f_i(x_i) + \sum_{j \neq i} \log f_{j|\mathsf{pa}_j}(x_j, x_{\mathsf{pa}_j}) \right) dx \\
&= h(g_i, f_i) + \sum_{j \neq i} \left(\int_{\mathbf{R}} g_{\mathsf{pa}_j}(\xi) h\left(g_{j|\mathsf{pa}_j}(\cdot, \xi), f_{j|\mathsf{pa}_j}(\cdot, \xi) \right) d\xi \right)
\end{aligned}
$$

- this problem *separates across dimension*
  - one problem to find f_i; a solution is $f_i = g_i$
  - since $g_{\mathsf{pa}_j} \geq 0$, minimize the integrand pointwise; a solution is $f_{j|\mathsf{pa}_j} = g_{j|\mathsf{pa}_j}$ for $j \neq i$

Mutual information

- we want to characterize the *optimal* tree, need notion of mutual information graph
- the *mutual information* of f_{ij} is $d_{kl}(f_{ij}, f_i f_j)$
  - we denote the symmetric *matrix of mutual informations* of f by $I(f)$, and define it by

$$I(g)_{ij} = d_{kl}(f_{ij}, f_i f_j)$$

- the *mutual information graph* of g is a weighted complete undirected graph on $\{1, \dots, d\}$
  - edge $\{i, j\}$ is weighted by $I(g)_{ij}$

Tree density approximator: second theorem

- **Theorem 2**: Let g be a density on $\mathbf{R}^d$. A tree T on $\{1, \ldots, d\}$ is an optimal approximator tree of g if and only if T is a maximum spanning tree of the mutual information graph of g.

Proof of theorem 2

- let f_T^* achieve minimum K-L divergence among densities which factor according to tree T
- we express the differential cross entropy of f_T^* relative to g as

$$
\begin{aligned}
h(g, f_T^*) &= h(g_1) - \sum_{j \neq 1} \left(\int_{\mathbf{R}^d} g(x) \log g_{j|\mathsf{pa}_j}(x_j, x_{\mathsf{pa}_j}) dx \right) \\
&= h(g_1) - \sum_{j \neq 1} \left(\int_{\mathbf{R}^d} g(x) \left(\log g_{j,\mathsf{pa}_j}(x_j, x_{\mathsf{pa}_j}) - \log g_{\mathsf{pa}_j}(x_{\mathsf{pa}_j}) \right) dx \right) \\
&= h(g_1) - \sum_{j \neq 1} \left(\int_{\mathbf{R}^d} g(x) \left(\log g_{j,\mathsf{pa}_j}(x_j, x_{\mathsf{pa}_j}) - \log g_{\mathsf{pa}_j}(x_{\mathsf{pa}_j}) - \log g_j(x_j) + \log g_j(x_j) \right) dx \right) \\
&= \sum_{i=1}^{d} h(g_i) - \sum_{j \neq 1} I(g)_{j,\mathsf{pa}_j} \\
&= \sum_{i=1}^{d} h(g_i) - \sum_{\{i,j\} \in T} I(g)_{ij}
\end{aligned}
$$

Approximating a normal

Normal densities

- $f : \mathbf{R}^d \to \mathbf{R}$ is a *normal density* on $\mathbf{R}^d$ if there exists $\Sigma \succ 0$ and μ such that

$$f(x) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^\mathsf{T}\Sigma^{-1}(x-\mu)\right)$$

  we write $f \sim \mathcal{N}(\mu, \Sigma)$; call $P = \Sigma^{-1}$ the *precision matrix*

- $f_i \sim \mathcal{N}(\mu_i, \Sigma_{ii})$ for $i = 1, \dots, d$

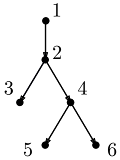
- $f_{ij} \sim \mathcal{N}\left(\begin{bmatrix}\mu_i \\ \mu_j\end{bmatrix}, \begin{bmatrix}\Sigma_{ii} & \Sigma_{ij} \\ \Sigma_{ji} & \Sigma_{jj}\end{bmatrix}\right)$ for $i, j = 1, \dots, d$ and $i \neq j$

- $f_{i|j}(\cdot, \xi) \sim \mathcal{N}\left(\mu_i + \Sigma_{ii}\Sigma_{jj}^{-1}\xi, \Sigma_{ii} - \Sigma_{ij}\Sigma_{jj}^{-1}\Sigma_{ji}\right)$ for $i, j = 1, \dots, d$ and $i \neq j$
  - define $\Sigma_{i|j}$ to be $\Sigma_{ii} - \Sigma_{ij}\Sigma_{jj}^{-1}\Sigma_{ji}$

- if f is a normal density, $I(f)_{ij} = -\frac{1}{2}\log(1 - \rho_{ij}^2)$, where $\rho_{ij} = \Sigma_{ij} / \left(\sqrt{\Sigma_{ii}\Sigma_{jj}}\right)$

Tree normal: example

- consider the same tree $T = \{\{1,2\},\{2,3\},\{2,4\},\{4,5\},\{4,6\}\}$ rooted at vertex 1



- if $f \sim \mathcal{N}(\mu, \Sigma)$ then
  - $f_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$
  - $f_{2|1}(\cdot, \xi) \sim \mathcal{N}\left(\Sigma_{21}\Sigma_{11}^{-1}\xi, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)$
  - $f_{3|2}(\cdot, \xi) \sim \mathcal{N}\left(\Sigma_{32}\Sigma_{22}^{-1}\xi, \Sigma_{33} - \Sigma_{32}\Sigma_{22}^{-1}\Sigma_{23}\right)$
  - $f_{4|2}(\cdot, \xi) \sim \mathcal{N}\left(\Sigma_{42}\Sigma_{22}^{-1}\xi, \Sigma_{44} - \Sigma_{42}\Sigma_{22}^{-1}\Sigma_{24}\right)$
  - $f_{5|4}(\cdot, \xi) \sim \mathcal{N}\left(\Sigma_{54}\Sigma_{44}^{-1}\xi, \Sigma_{55} - \Sigma_{54}\Sigma_{44}^{-1}\Sigma_{45}\right)$
  - $f_{6|4}(\cdot, \xi) \sim \mathcal{N}\left(\Sigma_{64}\Sigma_{44}^{-1}\xi, \Sigma_{66} - \Sigma_{64}\Sigma_{44}^{-1}\Sigma_{46}\right)$

Approximating a normal

- if g is normal, we have expressions for $g_{i|j}$
- **Theorem**: Let g be a normal density with mean $\mu \in \mathbf{R}^d$ and covariance $\Sigma \in \mathbf{S}^d_{++}$. Let T be an optimal approximator tree of g. Let f be a normal density with mean μ and precision matrix P where
  - $P_{11} = \Sigma_{11}^{-1} + \sum_{\mathsf{pa}_j=1} \Sigma_{j1}^2 \Sigma_{11}^{-2} \Sigma_{j|1}^{-1}$
  - for $i = 2, \ldots, d$, $P_{ii} = \Sigma_{i|\mathsf{pa}_i}^{-1} + \sum_{\mathsf{pa}_j=i} \Sigma_{ji}^2 \Sigma_{ii}^{-2} \Sigma_{j|i}^{-1}$
  - $i, j = 1, \ldots d$ and $i = \mathsf{pa}_j$, $P_{ij} = P_{ji} = -\Sigma_{ji} \Sigma_{jj}^{-1} \Sigma_{j|i}^{-1}$

  Then f is optimal tree approximator of g.
- f is *also a normal density*

Normal case proof

- use Theorem 1 to express f_T^* as

$$(1/c)\exp\left(-\frac{1}{2}\left(\Sigma_{11}^{-1}\bar{x}_1^2+\sum_{i\neq 1}\left(\bar{x}_i-\Sigma_{i,\mathsf{pa}_i}\Sigma_{\mathsf{pa}_i,\mathsf{pa}_i}^{-1}\bar{x}_{\mathsf{pa}_i}\right)^2\Sigma_{i|\mathsf{pa}_i}^{-1}\right)\right)$$

  with $\bar{x}_i = x_i - \mu_i$ and $c = \sqrt{(2\pi)^d\Sigma_{11}\prod_{i\neq 1}\Sigma_{i|\mathsf{pa}_i}}$.

- expand to express quadratic in the exponential as

$$\Sigma_{11}^{-1}\bar{x}_1^2+\sum_{i\neq 1}\left[\Sigma_{i|\mathsf{pa}_i}^{-1}\bar{x}_i^2-2\Sigma_{i,\mathsf{pa}_i}\Sigma_{\mathsf{pa}_i,\mathsf{pa}_i}^{-1}\Sigma_{i|\mathsf{pa}_i}^{-1}\bar{x}_i\bar{x}_{\mathsf{pa}_i}+\Sigma_{i,\mathsf{pa}_i}^2\Sigma_{\mathsf{pa}_i,\mathsf{pa}_i}^{-2}\Sigma_{i|\mathsf{pa}_i}^{-1}\bar{x}_{\mathsf{pa}_i}^2\right]$$

  so, with P defined as on previous slide $\bar{x}^\mathsf{T}P\bar{x}$ gives above

- c is $\sqrt{(2\pi)^d\det P^{-1}}$ since f_T^* integrates to one.

Example: the empirical normal

- let $x^1, \ldots, x^n \in \mathbf{R}^d$; the *empirical normal* density is a normal with
  - where $\bar{x} = \frac{1}{n} \sum_{k=1}^{n} x^k$ is the *empirical mean*
  - and $S = \frac{1}{n} \sum_{k=1}^{n} (x^k - \bar{x})(x^k - \bar{x})^{\mathsf{T}}$ is the *empirical covariance*
  - recall: these are the solutions to multivariate normal maximum likelihood density selection
- Theorem on previous slides gives following algorithm:

1. compute empirical mean and covariance of data
2. find maximum spanning tree of mutual information graph (edge weights are correlations)
3. take empirical mean, use empirical covariance to find the precision matrix

Maximum likelihood with tree normals

Maximum likelihood with tree normal

- have dataset $x^1, \ldots, x^n \in \mathbf{R}^d$ and want to find density f and tree T to

$$\begin{array}{ll} \text{maximize} & \displaystyle \frac{1}{n}\sum_{k=1}^{n} \log f(x^k) \\ \text{subject to} & f \text{ is normal and factors according to } T \end{array}$$

- **Theorem**: a normal density that factors according to a tree is a maximum likelihood density if and only if it is a optimal tree approximator of the empirical normal
  - on one hand, maximum likelihood leads us to approximating the empirical normal
  - on the other hand, our results about approximating normals solve the maximum likelihood problem

Proof of maximum likelihood equivalence

- let $g \sim \mathcal{N}(\mu_g, \Sigma_g)$, $f \sim \mathcal{N}(\mu_f, \Sigma_f)$ and $P_f = \Sigma_f^{-1}$.
- recall $d(g, f) = h(g, f) - h(g)$; second term constant w.r.t f, express first as:

$$h(g, f) = -\int_{\mathbf{R}^d} g \log f = \frac{1}{2}\left(\int_{\mathbf{R}^d} g \operatorname{tr}\big((x-\mu_f)^{\mathsf{T}}\Sigma_f^{-1}(x-\mu_f)\big) + \log\det\Sigma_f + \log(2\pi)^d\right)$$

  if $\mu_f = \mu_g$ then objective of approximation problem is equivalent to objective $\operatorname{tr}\Sigma_g P_f - \log\det P_f$

  - for optimizer, f matches g on one-variable marginals and so their means match
- likewise, express negative average log likelihood

$$-\frac{1}{n}\sum_{k=1}^{n}\log f(x^k) = \frac{1}{n}\frac{1}{2}\sum_{k=1}^{n}\operatorname{tr}\big((x-\mu_f)^{\mathsf{T}}\Sigma_f^{-1}(x-\mu_f)\big) + \log\det\Sigma_f + \log(2\pi)^d$$

- if μ_f is the empirical mean, then likelihood objective is equivalent to $\operatorname{tr} SP_f - \log\det P_f$
  - for optimizer, $\mu_f = \frac{1}{n}\sum_{k=1}^{n} x^k$ from first order conditions on objective

Takeaways

- we know how to approximate any density with one that factors according to a tree
- to be useful, need to know more about density, for example that it is normal
- picking the maximum likelihood tree normal is the same as approximating the empirical normal

Future work

- generalization to tree linear cascades
- use for Gaussian processes
- use in Kalman filters